Robust sparse canonical correlation analysis

Ines Wilms*and Christophe Croux

Leuven Statistics Research Centre (LStat), KU Leuven

Abstract

Background: Canonical correlation analysis (CCA) is a multivariate statistical method which describes the associations between two sets of variables. The objective is to find linear combinations of the variables in each data set having maximal correlation. In genomics, CCA has become increasingly important to estimate the associations between gene expression data and DNA copy number change data. The identification of such associations might help to increase our understanding of the development of diseases such as cancer. However, these data sets are typically high-dimensional, containing a lot of variables relative to the number of objects. Moreover, the data sets might contain atypical observations since it is likely that objects react differently to treatments. We discuss a method for Robust Sparse CCA, thereby providing a solution to both issues. Sparse estimation produces canonical vectors with some of their elements estimated as exactly zero. As such, their interpretability is improved. Robust methods can cope with atypical observations in the data.

Results: We illustrate the good performance of the Robust Sparse CCA method by several simulation studies and three biometric examples. Robust Sparse CCA considerably outperforms its main alternatives in (1) correctly detecting the main associations between the data sets, in (2) accurately estimating these associations, and in (3) detecting outliers.

Conclusions: Robust Sparse CCA delivers interpretable canonical vectors, while at the same time coping with outlying observations. The proposed method is able to describe the associations between high-dimensional data sets, which are nowadays commonplace in genomics. Furthermore, the Robust Sparse CCA method allows to characterize outliers.

Keywords

Canonical correlation analysis; penalized estimation; robust estimation

*Corresponding author ines.wilms@kuleuven.be

Background

Canonical correlation analysis (CCA), introduced by Hotelling (1936), identifies and quantifies the associations between two sets of variables. CCA searches for linear combinations, called *canonical variates*, of each of the two sets of variables having maximal correlation. The coefficients of these linear combinations are called the *canonical vectors*. The correlations between the canonical variates are called the *canonical correlations*. CCA is used to study associations in, for instance, genomic data (Tenenhaus et al., 2014), environmental data (Iaci et al., 2010), or biomedical data (Chen et al., 2013a). For more information on canonical correlations analysis, see e.g. Johnson and Wichern (1998), Chapter 10.

Sparse canonical vectors are canonical vectors with some of their elements estimated as exactly zero. The canonical variates then only depend on a subset of the variables, those corresponding to the non-zero elements of the estimated canonical vectors. Hence, the canonical variates are easier to interpret, in particular for high-dimensional data sets. Sparse estimation shows good performance in analyzing, for instance, genomic data (e.g. (Li et al., 2013; Fujita et al., 2007; Steinke et al., 2007)), or biological data (e.g. (Li and Ngom, 2013; August and Papachristodoulou, 2009)). Examples of CCA for high-dimensional data sets can be found in, for example, genetics (Gonzalez et al., 2008; Prabhakar and Fridley, 2012; Cruz-Cano and Lee, 2014) and machine learning (Sun et al., 2011).

Different approaches for sparse CCA have been proposed in the literature. Parkhomenko et al. (2009) use a sparse singular value decomposition to derive sparse singular vectors. Witten et al. (2009) develop a penalized matrix decomposition, and show how to apply it for sparse CCA. Waaijenborg et al. (2008); Lykou and Whittaker (2010); An et al. (2013); Wilms and Croux (2015) convert the CCA problem into a penalized regression framework to produce sparse canonical vectors. Chen et al. (2013b); Gao and Zhou (2014) discuss theoretical properties for sparse CCA. All these methods are not robust to outliers. A common problem in multivariate data sets, however, is the frequent occurrence of outliers. In genomics, for instance, some patients can react very differently to treatments because of their individual-specific genetic structure. Therefore, the possible presence of outlying observations should be taken into account.

Several *robust CCA* methods have been introduced in the literature. Dehon and Croux (2002) considers robust CCA using the Minimum Covariance Determinant estimator Rousseeuw and Van Driessen (1999). Asymptotic properties for CCA based on robust estimators of the covariance matrix are discussed in Taskinen et al. (2006). Branco et al. (2005) use a robust alternating regression approach to obtain the canonical variates. CCA can also be considered as a prediction problem, where the canonical variates obtained from the first data set serve as optimal predictors for the canonical variates of the second data set, and vice versa. As such, Adrover and Donato (2015) use a robust M-scale to evaluate the prediction quality, whereas Kudraszow and Maronna (2011) use a robust estimator of the multivariate linear model. None of these methods, however, are sparse.

This paper proposes a CCA method that is sparse and robust at the same time. As such, we deal with two important topics in applied statistics: sparse model estimation and the presence of outliers in the data. We use an alternating robust, sparse regression framework to sequentially obtain the canonical variates. Robust Sparse CCA has clear advantages: (i) it provides well interpretable canonical vectors since some of the elements of the canonical vectors are estimated as exactly zero, (ii) it is still computable for high-dimensional data sets, where the sample size exceeds the number of variables in each data set, (iii) it can cope with outliers in the data, which are even more likely to occur in high dimensions, and (iv) it provides

an interesting way to characterize these outliers.

Simulation studies were performed to investigate the performance of Robust Sparse CCA. These simulations show that Robust Sparse CCA achieves a substantially better performance compared to its leading alternatives CCA, Robust CCA and Sparse CCA. We illustrate the application of the Robust Sparse CCA method to an environmental data set and two genomic data sets. Robust Sparse CCA provides easy interpretable results. Moreover, we use Robust Sparse CCA to detect outlying observations in such high-dimensional data sets.

Methods

First, we consider the robust and sparse estimator for the CCA problem. Next, we discuss the algorithm. Finally, we discuss the simulation designs and performance measures used to compare the performance of Robust Sparse CCA to standard CCA, Robust CCA and Sparse CCA.

The estimator

We consider the CCA problem in a regression framework ((Brillinger, 1975; Izenman, 1975)). Given a sample of n observations $\mathbf{x}_i \in \mathbb{R}^p$ and $\mathbf{y}_i \in \mathbb{R}^q$ $(i = 1, \ldots, n)$. The two data matrices are denoted as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T$. We assume the data matrices are robustly centered using the median. The estimated canonical vectors are collected in the columns of the matrices $\widehat{\mathbf{A}} \in \mathbb{R}^{p \times r}$ and $\widehat{\mathbf{B}} \in \mathbb{R}^{q \times r}$. Here r is the number of canonical vectors. The columns of the matrices $\mathbf{X}\widehat{\mathbf{A}}$ and $\mathbf{Y}\widehat{\mathbf{B}}$ contain the estimates of the realizations of the canonical variates, and we denote their j^{th} column by $\hat{\mathbf{u}}_j$ and $\hat{\mathbf{v}}_j$, for $1 \leq j \leq r$. The objective function defining the canonical vector estimates is

$$(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) = \underset{(\mathbf{A},\mathbf{B})}{\operatorname{argmin}} \sum_{i=1}^{n} ||\mathbf{A}^T x_i - \mathbf{B}^T \mathbf{y}_i||^2. \tag{1}$$

The objective function in (1) is minimized under the restriction that each canonical variate $\hat{\mathbf{u}}_j$ is uncorrelated with the lower order canonical variates $\hat{\mathbf{u}}_k$, with $1 \leq k < j \leq r$. Similarly for the canonical vectors within the second set of variables. For identification purpose, a normalization condition requiring the canonical vectors to have unit norm is added. Typically, the canonical vectors are obtained by an eigenvalue analysis of a certain matrix involving the inverses of sample covariance matrices. But if $n < \max(p, q)$, these inverses do not exist.

We estimate the canonical vectors with an alternating regression procedure. If the matrix $\mathbf{A}$ in (1) is kept fixed, the matrix $\mathbf{B}$ can be obtained from a Least Squares regression of the canonical variates on $\mathbf{y}$ (and vice versa for estimating $\mathbf{A}$ keeping $\mathbf{B}$ fixed). The standard Least Squares estimator, however, is not sparse, nor robust to outliers. Therefore, we replace it by the sparse Least Trimmed Squares (sparse LTS) estimator (Alfons et al., 2013). The sparse LTS estimator can be applied to high-dimensional data and is robust to outliers.

The algorithm

We use a sequential algorithm to derive the canonical vectors.

First canonical vector pair. Denote the first canonical vector pair by $(\mathbf{A}_1, \mathbf{B}_1)$. Assume that the value of $\mathbf{A}_1$ is known. Denote the vector of squared residuals by $\mathbf{r}^2(\mathbf{B}_1) = (r_1^2, \ldots, r_n^2)^T$, with $r_i^2 = (\mathbf{A}_1^T\mathbf{x}_i - \mathbf{B}_1^T\mathbf{y}_i)^2, i = 1, \ldots, n$. The estimate of $\mathbf{B}_1$ is obtained as

$$\widehat{\mathbf{B}}_1|\mathbf{A}_1 = \underset{\mathbf{B}_1}{\operatorname{argmin}} \sum_{i=1}^{h} \left(\mathbf{r}^2(\mathbf{B}_1)\right)_{i:n} + h\lambda_{B_1} \sum_{j=1}^{q} |b_{1j}|, \tag{2}$$

where $\lambda_{B_1} > 0$ is a sparsity parameter, b_{1j} is the j^{th} element, $j = 1, \ldots, q$, of the first canonical vector $\mathbf{B}_1$, and $\left(\mathbf{r}^2(\mathbf{B}_1)\right)_{1:n} \leq \ldots \leq \left(\mathbf{r}^2(\mathbf{B}_1)\right)_{n:n}$ are the order statistics of the squared residuals. The canonical vector $\widehat{\mathbf{B}}_1$ is normed to length 1. The solution to (2) equals the sparse LTS estimator with $\mathbf{XA}_1$ as response and $\mathbf{Y}$ as predictor. Regularization by adding a penalty term to the objective function is necessary since the design matrix $\mathbf{Y}$ can be high-dimensional. Sparse model estimates are obtained by adding an L_1 penalty to the LTS objective function, similar as for the lasso regression estimator (Tibshirani, 1996). The sparse LTS estimator is computed with trimming proportion 25%, so size of the subsample $h = \lfloor 0.75n \rfloor$. To increase efficiency, we use a reweighting step. Further discussion and more detail on the sparse LTS estimator is provided in Additional file 1. As such, we get a robust sparse estimate $\widehat{\mathbf{B}}_1$.

Analogously, for a fixed value $\mathbf{B}_1$, denote the vector of squared residuals by $\mathbf{r}^2(\mathbf{A}_1) = (r_1^2, \ldots, r_n^2)^T$, with $r_i^2 = (\mathbf{B}_1^T\mathbf{y}_i - \mathbf{A}_1^T\mathbf{x}_i)^2, i = 1, \ldots, n$. The sparse LTS regression estimate of $\mathbf{A}_1$ with $\mathbf{YB}_1$ as response and $\mathbf{X}$ as predictor is given by

$$\widehat{\mathbf{A}}_1|\mathbf{B}_1 = \underset{\mathbf{A}_1}{\operatorname{argmin}} \sum_{i=1}^{h} \left(\mathbf{r}^2(\mathbf{A}_1)\right)_{i:n} + h\lambda_{A_1} \sum_{j=1}^{p} |a_{1j}|, \tag{3}$$

where $\lambda_{A_1} > 0$ is a sparsity parameter, a_{1j} is the j^{th} element, $j = 1, \ldots, p$ of the first canonical vector $\mathbf{A}_1$, and $\left(\mathbf{r}^2(\mathbf{A}_1)\right)_{1:n} \leq \ldots \leq \left(\mathbf{r}^2(\mathbf{A}_1)\right)_{n:n}$ are the order statistics of the squared residuals. The canonical vector $\widehat{\mathbf{A}}_1$ is normed to length 1.

This leads to an alternating regression scheme, updating in each step the estimates of the canonical vectors until convergence. After convergence of the algorithm, the values of $\mathbf{A}_1$ and $\mathbf{B}_1$ in subsequent iterations remain stable, and the same observations will be detected as outliers in regressions (2) and (3).

Higher order canonical vector pairs. We use deflated data matrices to estimate the higher order canonical vector pairs (see e.g. Branco et al. (2005)). For the second canonical vector pair, the deflated matrices are $\mathbf{X}_2^*$, the residuals of a column-by-column LTS regression of $\mathbf{X}$ on all lower order canonical variates, $\hat{\mathbf{u}}_1$ in this case; and $\mathbf{Y}_2^*$, the residuals of a column-by-column LTS regression of $\mathbf{Y}$ on $\hat{\mathbf{v}}_1$. Since these regressions only involve a small number of regressors, the standard LTS estimator with $\lambda = 0$ can be used.

The second canonical variate pair is then obtained by alternating between the following regressions until convergence:

$$\widehat{\mathbf{B}}_2^*|\mathbf{A}_2^* = \underset{\mathbf{B}_2^*}{\operatorname{argmin}} \sum_{i=1}^{h} \left(\mathbf{r}^2(\mathbf{B}_2^*)\right)_{i:n} + h\lambda_{B_2^*} \sum_{j=1}^{q} |b_{2j}^*|, \tag{4}$$

where $\mathbf{r}^2(\mathbf{B}_2^\star) = (r_1^2, \ldots, r_n^2)^T$, with $r_i^2 = (\mathbf{A}_2^{*T}\mathbf{x}_{2,i}^\star - \mathbf{B}_2^{\star T}\mathbf{y}_{2,i}^\star)^2, i = 1, \ldots, n$.

$$\widehat{\mathbf{A}}_2^*|\mathbf{B}_2^* = \underset{\mathbf{A}_2^*}{\operatorname{argmin}} \sum_{i=1}^{h} \left(\mathbf{r}^2(\mathbf{A}_2^*)\right)_{i:n} + h\lambda_{A_2^*} \sum_{j=1}^{p} |a_{2j}^*|, \tag{5}$$

where $\mathbf{r}^2(\mathbf{A}_2^\star) = (r_1^2, \ldots, r_n^2)^T$, with $r_i^2 = (\mathbf{B}_2^{*T}\mathbf{y}_{2,i}^\star - \mathbf{A}_2^{\star T}\mathbf{x}_{2,i}^\star)^2, i = 1, \ldots, n$. The canonical vectors $\widehat{\mathbf{B}}_2^*$ and $\widehat{\mathbf{A}}_2^*$ are both normed to length 1. We obtain $\hat{\mathbf{u}}_2^* = \mathbf{X}_2^*\widehat{\mathbf{A}}_2^*$ and $\hat{\mathbf{v}}_2^* = \mathbf{Y}_2^*\widehat{\mathbf{B}}_2^*$.

Finally, the second canonical vector needs to be expressed as linear combinations of the columns of the original data matrices, and not the deflated ones. Since we want to allow for zero coefficients in these linear combinations, a sparse approach is needed. To obtain a sparse $\widehat{\mathbf{A}}_2$, we regress $\hat{\mathbf{u}}_2^*$ on $\mathbf{X}$ using the sparse LTS estimator, yielding the fitted values $\hat{\mathbf{u}}_2 = \mathbf{X}\widehat{\mathbf{A}}_2$. To obtain a sparse $\widehat{\mathbf{B}}_2$, we regress $\hat{\mathbf{v}}_2^*$ on $\mathbf{Y}$ using the sparse LTS estimator, yielding the fitted values $\hat{\mathbf{v}}_2 = \mathbf{Y}\widehat{\mathbf{B}}_2$.

The higher order canonical variate pairs are obtained in a similar way. We perform alternating sparse LTS regressions as in (4) and (5), followed by a final sparse LTS step to retrieve the estimated canonical vectors $(\widehat{\mathbf{A}}_l, \widehat{\mathbf{B}}_l)$. It is not really necessary to use a sparse approach in regressions (4) and (5), other penalty functions can be used. A schematic representation of the complete algorithm is provided in Additional file 2.

Finally, note that as in other sparse CCA proposals (e.g. (Parkhomenko et al., 2009), (Witten et al., 2009), Waaijenborg et al. (2008), Wilms and Croux (2015)) the canonical variates are in general not uncorrelated. The robust sparse canonical vectors we obtain yield an interpretable basis of the space spanned by the canonical vectors. This basis can be made orthogonal (but not sparse) after suitable rotation if one desires so.

Initial value. A starting value for $\mathbf{A}_1$ is required to start up the algorithm. We compute the first robust principal component of $\mathbf{Y}$, denoted $\mathbf{z}_1$. The first robust principal component is calculated from the first eigenvector of the robustly estimated covariance matrix. For this aim, we use the spatial sign covariance estimator (Visuri et al., 2000). We regress $\mathbf{z}_1$ on $\mathbf{X}$ using the sparse LTS. The estimated regression coefficient matrix of this regression is used as initial value for $\mathbf{A}_1$. To obtain an initial estimate for the higher order canonical vectors $\mathbf{A}_l$, for $l = 2, \ldots, r$, we use the first robust principal component of the deflated data matrix and proceed analogously.

We performed several numerical experiments to investigate the sensitivity of the outcome of the algorithm to the choice of initial value. In low-dimensional settings, the choice of initial value is not important. In high-dimensional settings, a good initial value is more important. Note that the initial value should exist and be easily computable in all settings, which holds for our proposal.

Number of canonical variates to extract. To decide on the number of canonical variates r to extract, we use the maximum eigenvalue ratio criterion of An et al. (2013). We apply the Robust Sparse CCA algorithm and calculate the robust correlations $\hat{\rho}_1, \ldots, \hat{\rho}_{\text{rmax}}$, with $\text{rmax} = \min(p, q, 10)$. For high-dimensional data sets, we consider a maximum of 10 canonical correlations, since in practice, more than 10 canonical vector pairs are never used. Each $\hat{\rho}_j$ is obtained by computing the correlation between $\hat{\mathbf{v}}_j$ and $\hat{\mathbf{u}}_j$ from the bivariate Minimum Covariance Determinant (MCD) estimator with 25% trimming. Let $\hat{k}_j = \hat{\rho}_j / \hat{\rho}_{j+1}$ for $j = 1, \ldots, \text{rmax} - 1$. We extract r pairs of canonical variates, where $r = \text{argmax}_j \hat{k}_j$.

Convergence criterion. In each step of the alternating regression algorithm we update the estimates of the canonical vectors $\widehat{\mathbf{B}}_l^*$ and $\widehat{\mathbf{A}}_l^*$, for $l = 1, \ldots, r$. We iterate until the relative change in the value of the convergence criterion in two successive iterations[1] is smaller than the convergence tolerance value $\epsilon = 10^{-2}$. As convergence criterion, we consider

$$\text{Convergence criterion} = \frac{1}{h} \sum_{i=1}^{h} (\mathbf{r}^2(\widehat{\mathbf{A}}_l^*, \widehat{\mathbf{B}}_l^*))_{i:n},$$

[1] One iteration includes one cycle of estimating $\mathbf{A}_l^* | \mathbf{B}_l^*$ and $\mathbf{B}_l^* | \mathbf{A}_l^*$.

for $l = 1, \ldots, r$, where $\mathbf{r}^2(\widehat{\mathbf{A}}_l^*, \widehat{\mathbf{B}}_l^*) = (r_1^2, \ldots, r_n^2)^T$, with $r_i^2 = (\widehat{\mathbf{A}}_l^{*T}\mathbf{x}_{l,i}^\star - \widehat{\mathbf{B}}_l^{*T}\mathbf{y}_{l,i}^\star)^2, i = 1, \ldots, n$. $\mathbf{X}_l^*$ and $\mathbf{Y}_l^*$ are the original data sets for $l = 1$, and the deflated data matrices for $l = 2, \ldots, r$. In the simulations we conducted, convergence was almost always reached.[2] For data sets with $n = 100, p = q = 10$, on average 6 iterations per canonical vector pair are needed to converge. For $n = 50, p = q = 100$, on average 10 iterations are needed to converge.

Choice of the sparsity parameter. The sparsity parameters controlling the penalization on the regression coefficient matrices are selected with the Bayesian Information Criterion (e.g. (Yin and Li, 2011)). We use a range of values for the sparsity parameters and select the one with the lowest value of

$$\text{BIC}_{\lambda_{\widehat{\mathbf{A}}_l^*}} = n \cdot \log\left(\frac{1}{h}\sum_{i=1}^{h}\left(\mathbf{r}^2(\widehat{\mathbf{A}}_l^*)\right)_{i:n}\right) + df_{\lambda_{\widehat{\mathbf{A}}_l^*}} \cdot \log(n),$$

$$\text{BIC}_{\lambda_{\widehat{\mathbf{B}}_l^*}} = n \cdot \log\left(\frac{1}{h}\sum_{i=1}^{h}\left(\mathbf{r}^2(\widehat{\mathbf{B}}_l^*)\right)_{i:n}\right) + df_{\lambda_{\widehat{\mathbf{B}}_l^*}} \cdot \log(n),$$

for $l = 1, \ldots, r$, with $df_{\lambda_{\widehat{\mathbf{A}}_l^*}}$ and $df_{\lambda_{\widehat{\mathbf{B}}_l^*}}$ the respective number of non-zero estimated regression coefficients.

Computation time. All computations are carried out in `R` version 3.2.1. The code of the algorithm is made available on a webpage of the first author (http://feb.kuleuven.be/ines.wilms/software). For data sets with $n = 100, p = q = 10$, on average 10 seconds are needed to extract one canonical vector pair on an Intel Xeon E5-2699 v3 @ 2.30GHz machine. For $n = 50, p = q = 100$, we need 540 seconds on average, for $n = 100, p = q = 10000$, computation time increases to 11 hours on average. But even in high dimensions, the number of iterations remains lows (8 on average). The high computing time needed for $p = q = 10000$ is mainly due to the sparse LTS estimator, taken from the `R`-package `robustHD` (Alfons, 2014). By including a variable screening step Fan and Lv (2008) preceding the computation of the sparse LTS estimator, one could reduce the total computation time considerably.

Simulation designs

To investigate the performance of Robust Sparse CCA, we conduct a simulation study. We consider several simulation designs.

In the "Uncorrelated Sparse Low-dimensional" and "Correlated Sparse Low-dimensional" design, there is one canonical variate pair and the canonical vectors have a sparse structure. The variables within each data set are uncorrelated in the first design, and correlated in the second design. In the "NonSparse Low-dimensional" design, there are two canonical variate pairs and the canonical vectors are non-sparse. The remaining three designs are high-dimensional with a lot of variables compared to the sample size. Only Sparse CCA and Robust Sparse CCA can be computed. In the "Sparse High-dimensional 1" design with $n = 100, p = 100, q = 4$, there is one canonical variate pair and the canonical vectors are sparse. In the "Sparse High-dimensional 2" design with $n = 100, p = q = 100$, there is one canonical variate pair and each canonical vector contains ten non-zero elements. In the "Sparse Ultra High-dimensional" design there are much more variables ($p = q = 10000$) than observations ($n = 100$). There is one canonical variate pair and each canonical vector contains ten non-zero elements. The number of simulations for each design except the

[2]Less than 5% of all simulation runs did not reach convergence after 50 iterations. In case of non-convergence, results from the last iteration run are taken.

last one is $M = 1000$. For the "Sparse Ultra High-dimensional design" $M = 100$ to reduce computational burden.

For each design, the following settings are considered

(a) *No contamination.* We generate data matrices $\mathbf{X}$ and $\mathbf{Y}$ according to a multivariate normal distribution $N_{p+q}(\mathbf{0}, \mathbf{\Sigma})$, with covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{xx} & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{xy}^T & \mathbf{\Sigma}_{yy} \end{bmatrix}$$

described in Table 1.

(b) *t-distribution.* We generate data matrices $\mathbf{X}$ and $\mathbf{Y}$ according to a multivariate t-distribution with three degrees of freedom $t_3(\mathbf{0}, \mathbf{\Sigma})$.

(c) *Contamination.* 90% of the data are generated from $N_{p+q}(\mathbf{0}, \mathbf{\Sigma})$, and 10% of the data are generated from $N_{p+q}(\mathbf{2}, \mathbf{\Sigma}_{\text{cont}})$, with

$$\mathbf{\Sigma}_{\text{cont}} = \begin{bmatrix} \mathbf{\Sigma}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{yy} \end{bmatrix}.$$

Similar conclusions can be drawn from other contamination settings (e.g. where only one of the two data sets is contaminated) and are available from the authors upon request.

Performance measures

In our simulation study, the estimators are evaluated on their estimation accuracy and sparsity recognition performance.

For evaluating estimation accuracy, we compute for each simulation run m, with $m = 1, \ldots, M$, the angle $\theta^m(\hat{\mathbf{A}}^m, \mathbf{A})$ between the subspace spanned by the estimated canonical vectors (contained in the columns of $\hat{\mathbf{A}}^m$) and the subspace spanned by the true canonical vectors (contained in the columns of $\mathbf{A}$). We proceed analogously for the matrix $\mathbf{B}$. The average angles, measuring the estimation accuracy, are given by

$$\bar{\theta}(\hat{\mathbf{A}}, \mathbf{A}) = \frac{1}{M} \sum_{m=1}^{M} \theta^m(\hat{\mathbf{A}}^m, \mathbf{A}) \quad \text{and} \quad \bar{\theta}(\hat{\mathbf{B}}, \mathbf{B}) = \frac{1}{M} \sum_{m=1}^{M} \theta^m(\hat{\mathbf{B}}^m, \mathbf{B}).$$

For evaluating sparsity, we use the true positive rate and the true negative rate

$$\text{TPR}(\hat{\mathbf{A}}^m, \mathbf{A}) = \frac{\#\{(i,j) : \widehat{\mathbf{A}}_{ij}^m \neq 0 \ \text{ and } \ \mathbf{A}_{ij} \neq 0\}}{\#\{(i,j) : \mathbf{A}_{ij} \neq 0\}}$$
$$\text{TNR}(\hat{\mathbf{A}}^m, \mathbf{A}) = \frac{\#\{(i,j) : \widehat{\mathbf{A}}_{ij}^m = 0 \ \text{ and } \ \mathbf{A}_{ij} = 0\}}{\#\{(i,j) : \mathbf{A}_{ij} = 0\}}.$$

We proceed analogously for the matrix $\mathbf{B}$. A true positive is a coefficient that is non-zero in the true model, and is estimated as non-zero. A true negative is a coefficient that is zero in the true model, and is estimated as zero. Both should be as high as possible for a sparse estimator. Note that the false positive rate is the complement of the true negative rate (i.e. FPR=1-TNR). A sparse estimator should control the FPR, which can be seen as a false discovery rate, at a sufficiently low level.

In our empirical applications, to decide on the number of canonical variate pairs to extract, we use the maximum eigenvalue ratio criterion, as discussed in the "Methods" Section. To compare the performance of

the CCA approaches, we perform a leave-one-out cross-validation exercise and compute the cross-validation score

$$CV = \frac{1}{r}\frac{1}{h}\sum_{i=1}^{h} ||\widehat{\mathbf{A}}_{-i}^{T}\mathbf{x}_i - \widehat{\mathbf{B}}_{-i}^{T}\mathbf{y}_i||^2, \tag{6}$$

where $\widehat{\mathbf{A}}_{-i}^{T}$ and $\widehat{\mathbf{B}}_{-i}^{T}$ contain the estimated canonical vectors when the i^{th} observation is left out of the estimation sample and $h = \lfloor n(1-\alpha) \rfloor$, with $\alpha = 0$ (0% Trimming) or $\alpha = 0.1$ (10% Trimming). We use trimming to eliminate the effect of outliers in the cross-validation score.

Results

Simulation Study

We compare the performance of the Robust Sparse CCA method with (i) standard CCA, (ii) Robust CCA, and (iii) Sparse CCA. The alternating regression algorithm is used for all four estimators, for ease of comparability. Robust CCA uses LTS instead of sparse LTS, and corresponds to the alternating regression approach of Branco et al. (2005). Sparse CCA uses the lasso instead of sparse LTS, Pearson correlations for computing the canonical correlations, and ordinary PCA for getting the initial values. The sparsity parameters for sparse CCA are selected with BIC. Standard CCA is like sparse CCA, but using the LS instead of the lasso.

Summary results for the estimator $\widehat{\mathbf{A}}$ are in Table 2. The results for $\widehat{\mathbf{B}}$ are similar and, therefore, omitted. Standard errors around the average angles, TPRs and TNRs are in almost all cases smaller than 6% of the reported numbers in Table 2.

First we discuss the results from the "Uncorrelated Sparse Low-dimensional" design. In the scenario without contamination, the sparse estimators Sparse CCA and Robust Sparse CCA achieve a much better average estimation accuracy than the non-sparse estimators CCA and Robust CCA. As expected, a sparse method results in increased estimation accuracy when the true canonical vectors have a sparse structure. Looking at sparsity recognition performance, Sparse CCA and Robust Sparse CCA perform equally good in retrieving the sparsity in the data generating process. In the contaminated simulation setting, the robust estimators maintain their accuracy. Robust Sparse CCA performs best and clearly outperforms Robust CCA: for instance, Robust Sparse CCA achieves an average estimation accuracy of 0.05 against 0.15 for the contamination setting, see Table 2. The non-robust estimators CCA and Sparse CCA are clearly influenced by the outliers, as reflected by the much higher values of the average angle $\bar{\theta}(\hat{\mathbf{A}}, \mathbf{A})$ in Table 2. Sparse CCA now performs even worse than Robust CCA. The considered contamination induces overfitting in Sparse CCA, reflected in the low values of the true negative rate, or alternatively, the high values of the false positive rate.

In an unreported simulation study, we investigated the effect of the signal strength on the results. We vary the value of the true canonical correlation in the first design from 0.1 to 0.9, thereby increasing the signal strength. If outliers are present, Robust Sparse CCA always performs best. The margin by which it outperforms Sparse CCA is larger if the signal is stronger. If no outliers are present, Sparse CCA performs best for weak signal levels below 0.6.

Similar conclusions can be drawn from the "Correlated Sparse Low-dimensional" and "NonSparse Low-dimensional" design. Note that the true negative rate in Table 2 is omitted for the "NonSparse Low-

dimensional" design since the true canonical vectors are non-sparse. In the situation without contamination, the price the sparse methods pay in the "NonSparse Low-dimensional" design is a decreased estimation accuracy, as measured by the average angle. For Robust Sparse CCA compared to Robust CCA this decrease is marginal. In the contaminated settings, the robust methods perform best and show similar performance.

For the high-dimensional designs, only Sparse CCA and Robust Sparse CCA are computable. For the "Sparse High-dimensional 1" design, Robust Sparse CCA is competitive to Sparse CCA if no outliers are present. When adding outliers, the performance of Sparse CCA gets distorted. For the heavier tailed t-distribution, the average estimation accuracy of Robust Sparse CCA compared to Sparse CCA is much better: 0.56 against 0.70. For the contamination setting, the average estimation accuracy of Robust Sparse CCA is even more than twice as good as the average estimation accuracy of Sparse CCA. Similar conclusions hold for the second high-dimensional design.

In the "Sparse Ultra High-dimensional" design, Sparse CCA performs best if no outliers are present. For the heavier tailed $t-$distribution, Robust Sparse CCA and Sparse CCA perform comparable in terms of estimation accuracy. But in the presence of outliers, Robust Sparse CCA improves estimation accuracy of Sparse CCA by about 22%. Moreover, Robust Sparse CCA achieves a good balance between the TPR and the TNR, while Sparse CCA suffers from a low TPR if outliers are present.

In sum, Robust Sparse CCA shows the best overall performance in this simulation study. It performs best in sparse contaminated settings. In sparse non-contaminated settings, Robust Sparse CCA is competitive to Sparse CCA. In contaminated non-sparse settings, Robust Sparse CCA is competitive to Robust CCA.

Comparison of Robust Sparse CCA to other CCA alternatives

We compare the performance of Robust Sparse CCA to

- the sparse CCA methods of Parkhomenko et al. (2009), Witten et al. (2009), and Waaijenborg et al. (2008). The sparsity parameters of all methods are selected as proposed by the respective authors. Note that these methods are not robust.

- sparse CCA applied on pre-processed data. As a pre-processing step to remove outliers, we transformed the data towards normality by replacing them by their normal scores (see e.g. (Rousseeuw and Leroy, 1987), page 150).

- sparse CCA using the robust initial value for the algorithm as Robust Sparse CCA.

Summary results for the estimator $\widehat{\mathbf{A}}$ are in Table 3. For reasons of brevity, we only report the results from the "Sparse High-dimensional 2" design. Similar conclusions are obtained from the other designs and are available from the authors upon request.

If no outliers are present, (i) Robust Sparse CCA is competitive to the sparse CCA methods of Parkhomenko et al. (2009), Witten et al. (2009), and Waaijenborg et al. (2008). (ii) Robust Sparse CCA performs comparable to Sparse CCA on pre-processed data. (iii) Sparse CCA with the same initial value as Robust Sparse CCA performs comparable to Sparse CCA.

If outliers are present, (i) Robust Sparse CCA outperforms the sparse CCA methods of Parkhomenko et al. (2009), Witten et al. (2009), and Waaijenborg et al. (2008). (ii) Robust Sparse CCA outperforms Sparse CCA on pre-processed data. Sparse CCA on pre-processed data performs better than Sparse CCA. (iii)

Robust Sparse CCA outperforms Sparse CCA with the same initial value. Here, differences in performance between Robust Sparse CCA and Sparse CCA stem from the use of the sparse LTS instead of the lasso regressions. Hence, the use of the sparse LTS estimator in the alternating regression scheme is essential.

Applications

We consider three biometric applications. The first data set is low-dimensional and often used in Robust Statistics. The other two data sets are high-dimensional and have been used before in papers on sparse CCA. We show that the performance of Robust Sparse CCA on these data sets is much better than the performance of Sparse CCA.

Evaporation data set

We analyze an environmental data set from Freund (1979). Two sets of environmental variables have been measured on $n = 46$ consecutive days from June 6 until July 21.[3] The first set contains $p = 3$ soil temperature variables (maximum, minimum and average soil temperature). The second set contains $q = 7$ environmental variables (maximum, minimum and average air temperature; maximum, minimum and average daily relative humidity; and total wind). The aim is to find and quantify the relations between the soil temperature variables and the remaining variables.

As a first inspection of the data, we use the Distance-Distance plot (Rousseeuw and van Zomeren, 1990) in Figure 1. The Distance-Distance plot displays the robust distances versus the Mahalanobis distances. The vertical and horizontal lines are drawn at values equal to the square root of the 97.5% quantile of a chi-squared distribution with 10 degrees of freedom. Points beyond those lines would be considered as outliers. The Distance-Distance plot reveals some outliers: objects 31 and 32, for example, are extreme outliers. This suggests the need for a robust CCA method. Table 4 reports the cross-validation scores from equation (6) for the four CCA methods. For all methods two canonical variate pairs are extracted. The method that achieves the lowest cross-validation score has the best out-of-sample performance. Robust Sparse CCA achieves the best cross-validation score.

Table 5 shows the estimated canonical vectors for the Robust CCA and Robust Sparse CCA method. By adding the penalty term, the number of non-zero coefficients in the two canonical vectors is reduced from a total of 20 for Robust CCA to 10 for Robust Sparse CCA. The price to pay for the sparseness is a slight decrease in the estimated canonical correlations (computed using the bivariate MCD estimator, see "Methods" Section): they drop from 0.93 to 0.87 for the first one, and from 0.56 to 0.48 for the second canonical correlation. We find this decrease acceptable, given the gained sparsity in the canonical vectors. The sparse structure of the canonical vectors facilitates interpretation. The first canonical variate in the soil temperature data set, for instance, is uniquely determined by the variable AVST.

Nutrimouse data set

This genetic data set is publicly available in the R package `CCA` (Gonzalez et al., 2008). Two sets of variables, i.e. gene expressions and fatty acids, are available for $n = 40$ mice. The first set contains expressions of $p = 120$ genes measured in liver cells. The second set of variables contains concentrations of $q = 21$ hepatic

[3] We treat the different measurements from the consecutive days as being independent from each other.

fatty acids (FA). In this experiment, there are two groups of mice (wild-type and PPARα deficient mice) that receive a specific diet (five possible diets). More details on how the data were obtained can be found in Martin et al. (2007). The aim is to identify a small set of genes that are correlated with the fatty acids.

In this data set, the number of experimental units is smaller than the number of variables. Therefore, standard CCA nor Robust CCA can be performed. Robust Sparse CCA and Sparse CCA can be applied in this high-dimensional setting and produce interpretable, sparse canonical vectors. For both methods, one canonical variate pair is extracted. The cross-validation scores from equation (6) are reported in Table 6. Robust Sparse CCA outperforms Sparse CCA. The cross-validation scores are reduced by about 90% when using the robust method.

Given its better out-of-sample performance, we discuss the estimated canonical vectors obtained using Robust Sparse CCA. The top panel of Figure 2 displays the coefficients of the selected genes, i.e. those genes with non-zero estimated coefficients, in the first canonical vector: 24 out of 120 variables are selected. The solution is very sparse, facilitating interpretation. Martin et al. (2007) find a consistent reduction of `Cyp3a11` in PPARα livers on the one hand, and an overexpression of `CAR1` on the other hand. Both genes are selected and have among the highest (absolute) coefficients. The coefficients of the selected fatty acids are displayed in the bottom panel of Figure 2: 13 out of 21 fatty acid variables are selected. The fatty acids `C22:6n-3`, `C22:5n-3`, `C22:5n-6`, `C22:4n-3` and `C20:5n-3` are related to the effect of the five diets used in this experiment. From Figure 2, we see that four out of these five fatty acids are selected.

Breast cancer data set

The genetic data set is described in Chin et al. (2006) and available in the `R package PMA` (Witten et al., 2011). Two sets of data, i.e. gene expression data (19 672 variables) and comparative genomic hybridization (CGH) data (2149 variables) are available for $n = 89$ patients, and this for 23 chromosomes. We analyze the data for each of the chromosomes separately, each time using the CGH and gene expression variables for that particular chromosome. Depending on the chromosome, either 1, 2, 3, or 4 canonical vector pairs are extracted. The aim is to identify a subset of CGH variables that are correlated with a subset of gene expression variables.

Results of the cross-validation scores of equation (6) are reported in Figure 3. For each of the 23 chromosomes, we plot the value of the cross-validation score (0% trimming) for Robust Sparse CCA (horizontal axis) and Sparse CCA (vertical axis). Results when using 10% trimming are similar and, therefore, omitted. The cross-validation scores of Robust Sparse CCA are much better than those of Sparse CCA: all points are lying above the 45°-line. For chromosomes 1, 3, 4, and 11, for instance, the cross-validation scores of Robust Sparse CCA are more than 10 times lower than those of Sparse CCA. Since Robust Sparse CCA performs much better, outliers might be present for these chromosomes. Hence, it is safer to use Robust Sparse CCA instead of Sparse CCA.

The Robust Sparse CCA method yields an interesting way to characterize the outliers. To this end, we create the Residual Distance plot of the residuals $\mathbf{X}\widehat{\mathbf{A}} - \mathbf{Y}\widehat{\mathbf{B}}$, and this for each of the 23 chromosomes. The Residual Distance plot displays the robust distance of the residuals (vertical axis) versus the observation number (horizontal axis). Points above the horizontal black line are marked as outliers. Results for chromosome 3 and 8 are displayed in Figure 4, results for the other chromosomes are available upon request. For some chromosomes, like chromosome 3, the difference in cross-validation scores of Robust Sparse CCA

and Sparse CCA in Figure 3 is outspoken, suggesting that outliers might be present. We use the Residual Distance plot (Figure 4, left panel) to detect which patients are outlying. In the Residual Distance plot of chromosome 3 a lot of patients are marked as outliers. For chromosome 8, on the other hand, the cross-validation scores of Sparse CCA and Robust Sparse CCA are nearly identical, which might suggest that there are no outliers. Looking at the Residual Distance Plot of chromosome 8 (Figure 4, right panel), no outliers are indeed detected.

Discussion

Robust Sparse CCA has three important advantages over Robust CCA. (i) Robust Sparse CCA improves model interpretation since only a limited number of variables, those corresponding to the non-zero elements of the canonical vectors, enter the estimated canonical variates (cfr. evaporation application), (ii) if the number of variables approaches the sample size, the estimation precision of Robust CCA suffers, and (iii) if the number of variables exceeds the sample size, Robust CCA can not even be performed. Robust Sparse CCA can still be computed (cfr. nutrimouse and breast cancer application).

The key ingredient of the Robust Sparse CCA algorithm is the sparse LTS proposed by Alfons et al. (2013). The choice of the subsample size h, see equation (2) involves a trade-off between robustness and estimation accuracy. We use $h = \lfloor 0.75 \cdot n \rfloor$, as recommended by Alfons et al. (2013). This guarantees a sufficiently high estimation accuracy and a good robustness/accuracy trade-off. If the researcher thinks that the proportion of outliers in one of the two data sets is larger than 25%, one could consider higher values of h. Our Robust Sparse CCA algorithm starts by robustly centering each variable using the coordinatewise median. The spatial median (e.g. Rousseeuw and Leroy (1987), page 251) could serve as an alternative to the coordinatewise median.

Several questions are left for future research. One could use a joint selection criterion for the number of canonical variate pairs and the sparsity parameter. This would, however, increase computation time substantially. To obtain sparse canonical vectors, we use a Lasso penalty. Other penalty functions such as the Adaptive Lasso (Zou, 2006) could be considered. The Adaptive Lasso is consistent for variable selection, whereas the Lasso is not. Furthermore, we use a regularized version of the LTS estimator. One could also use a regularized version of the S-estimator or the MM-estimator to increase efficiency. Up to our knowledge, however, the sparse LTS is the only robust sparse regression estimator for which efficient code (Alfons, 2014) is available.

Conclusion

Sparse Canonical Correlation Analysis delivers interpretable canonical vectors, with some of its elements estimated as exactly zero. Robust Sparse CCA retains this advantage, while at the same time coping with outlying observations.

Typically, the canonical vectors are based on the sample versions of the covariance matrices. One could think of estimating those covariance matrices with an estimator that is robust and sparse at the same time, and then, to compute the eigenvectors. This approach would result in canonical vectors being robust, however, not sparse. To circumvent this pitfall, we reformulate the CCA problem in a regression framework.

Nowadays, high-dimensional data sets where the researcher suspects contamination to be present are commonplace in genetics. This requires tailored methods such as Robust Sparse CCA to analyze the information they contain.

References

Adrover, J. and Donato, S. (2015), "A robust predictive approach for canonical correlation analysis," *Journal of Multivariate Analysis*, 133, 356–376.

Alfons, A. (2014), *robustHD: Robust methods for high-dimensional data*, R package version 0.5.0.

Alfons, A.; Croux, C. and Gelper, S. (2013), "Sparse least trimmed squares regression for analyzing high-dimensional large data sets," *The Annals of Applied Statistics*, 7(1), 226–248.

An, B.; Guo, J. and Wang, H. (2013), "Multivariate regression shrinkage and selection by canonical correlation analysis," *Computational Statistics and Data Analysis*, 62, 93–107.

August, E. and Papachristodoulou, A. (2009), "Efficient, sparse biological network determination," *BMC Systems Biology*, 3:25.

Branco, J.; Croux, C.; Filzmoser, P. and Oliviera, M. (2005), "Robust Canonical Correlations: A Comparative Study," *Computational Statistics*, 20, 203–229.

Brillinger, D. (1975), *Time Series: Data analysis and theory*, New York: Holt, Rinehart, and Winston.

Chen, J.; Bushman, F.; Lewis, J.; Wu, G. and Li, H. (2013a), "Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis," *Biostatistics*, 14(2), 244–258.

Chen, M.; Gao, C.; Ren, Z. and Zhou, H. (2013b), "Sparse CCA via precision adjusted iterative thresholding," *arXiv:1311.6186*.

Chin, K.; DeVries, S.; Fridlyand, J.; Spellman, P.; Roydasgupta, R.; Kuo, W.; Lapuk, A.; Neve, R.; Qian, Z.; Ryder, T. et al. (2006), "Genomic and transcriptional aberrations linked to breast cancer pathophysiologies," *Cancer Cell*, 10(6), 529–541.

Cruz-Cano, R. and Lee, M.-L. (2014), "Fast regularized canonical correlation analysis," *Computational Statistics and Data Analysis*, 70, 88–100.

Dehon, C. and Croux, C. (2002), "Analyse canonique basée sur des estimateurs robustes de la matrice de covariance," *La Revue de Statistique Appliquée*, 2, 5–26.

Fan, J. and Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society Series B*, 70(5), 849–911.

Freund, R. (1979), "Multicollinearity etc. Some 'new' examples," *American Statistical Association Proceedings of Statistical Computing Section*, 111–112.

Fujita, A.; Sato, J.; Garay-Malpartida, H.; Yamaguchi, R.; Miyano, S.; Sogayar, M. and Ferreira, C. (2007), "Modeling gene expression regulatory networks with the sparse vector autoregressive model," *BMC Systems Biology*, 1:39.

Gao, C., M. Z. and Zhou, H. (2014), "Sparse CCA: adaptive estimation and computational barriers," *arXiv:1409.8565*.

Gonzalez, I.; Dejean, S.; Martin, P. and Baccini, A. (2008), "CCA: An R package to extend canonical correlation analysis," *Journal of Statistical Software*, 23(12), 1–14.

Hotelling, H. (1936), "Relations between two sets of variates," *Biometrika*, 28, 321–377.

Iaci, R.; Sriram, T. N. and Yin, X. (2010), "Multivariate association and dimension reduction: A generalization of canonical correlation analysis," *Biometrics*, 66(4), 1107–1118.

Izenman, A. (1975), "Reduced-rank regression for the multivariate linear model," *Journal of Multivariate Analysis*, 5(2), 248–264.

Johnson, R. and Wichern, D. (1998), *Applied Multivariate Statistical Analysis*, London: Prentice-Hall.

Kudraszow, N. and Maronna, R. (2011), "Robust canonical correlation analysis: a predictive approach," *Working paper*.

Li, J.; Lin, D.; Cao, H. and Wang, Y. (2013), "An improved sparse representation model with structural information for Multicolour Fluorescence In-Situ Hybridization (M-FISH) image classification," *BMC Systems Biology*, 7(4):S5.

Li, Y. and Ngom, A. (2013), "Sparse representation approaches for the classification of high-dimensional biological data," *BMC Systems Biology*, 7(4):S6.

Lykou, A. and Whittaker, J. (2010), "Sparse CCA using a lasso with positivity constraints," *Computational Statistics and Data Analysis*, 54(12), 3144–3157.

Martin, P.; Guillon, H.; Lasserre, F.; Dejean, S.; Lan, A.; Pascussi, J.; SanCristobal, M.; Legrand, P.; Besse, P. and Pineau, T. (2007), "Novel aspects of PPARα-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study," *Hepatology*, 45(3), 767–777.

Parkhomenko, E.; Tritchler, D. and Beyene, J. (2009), "Sparse canonical correlation analysis with application to genomic data integration," *Statistical Applications in Genetics and Molecular Biology*, 8(1), 1–34.

Prabhakar, C. and Fridley, B. (2012), "Comparison of penalty functions for sparse canonical correlation analysis," *Computational Statistics and Data Analysis*, 56(2), 245–254.

Rousseeuw, P. and Leroy, A. (1987), *Robust regression and outlier detection*, New York: John Wiley & Sons.

Rousseeuw, P. and Van Driessen, K. (1999), "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, 41(3), 212–223.

Rousseeuw, P. and van Zomeren, B. (1990), "Unmasking multivariate outliers and leverage points," *Journal of the American Statistical Association*, 85(411), 633–639.

Steinke, F.; Seeger, M. and Tsuda, K. (2007), "Experimental design for efficient identification of gene regulatory networks using sparse Bayesian models," *BMC Systems Biology*, 1:51.

Sun, L.; Ji, S. and Ye, J. (2011), "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1), 194–200.

Taskinen, S.; Croux, C.; Kankainen, A.; Ollila, E. and Oja, H. (2006), "Canonical Analysis based on Scatter Matrices," *Journal of Multivariate Analysis*, 97, 359–384.

Tenenhaus, A.; Philippe, C.; Guillemot, V.; Le Cao, K.; Grill, J. and Frouin, V. (2014), "Variable selection for generalized canonical correlation analysis," *Biostatistics*, 15(3), 569–583.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B*, 58(1), 267–288.

Visuri, S.; Koivunen, V. and Oja, H. (2000), "Sign and rank covariance matrices," *Journal of Statistical Planning and Inference*, 91(2), 557–575.

Waaijenborg, S.; Hamer, P. and Zwinderman, A. (2008), "Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis," *Statistical Applications in Genetics and Molecular Biology*, 7(1), Article 3.

Wilms, I. and Croux, C. (2015), "Sparse canonical correlation analysis from a predictive point of view," *Biometrical Journal*, 57(5), 834–851.

Witten, D.; Tibshirani, R. and Gross, S. (2011), *Penalized multivariate analysis*, R package version 1.0.7.1.

Witten, D.; Tibshirani, R. and Hastie, T. (2009), "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, 10(3), 515–534.

Yin, J. and Li, H. (2011), "A sparse conditional gaussian graphical model for analysis of genetical genomics data," *The Annals of Applied Statistics*, 5(4), 2630–2650.

Zou, H. (2006), "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101(476), 1418–1429.

Figures

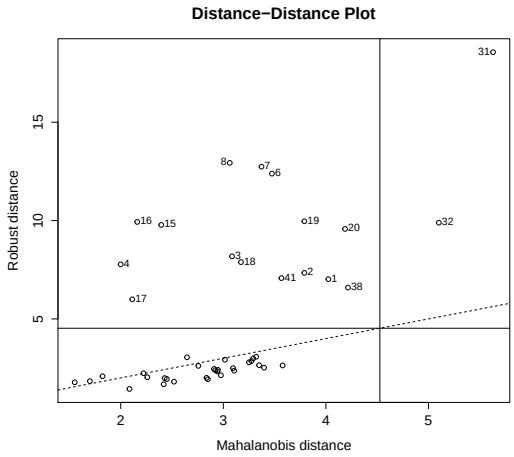

Figure 1: Evaporation data set: Distance-Distance Plot.

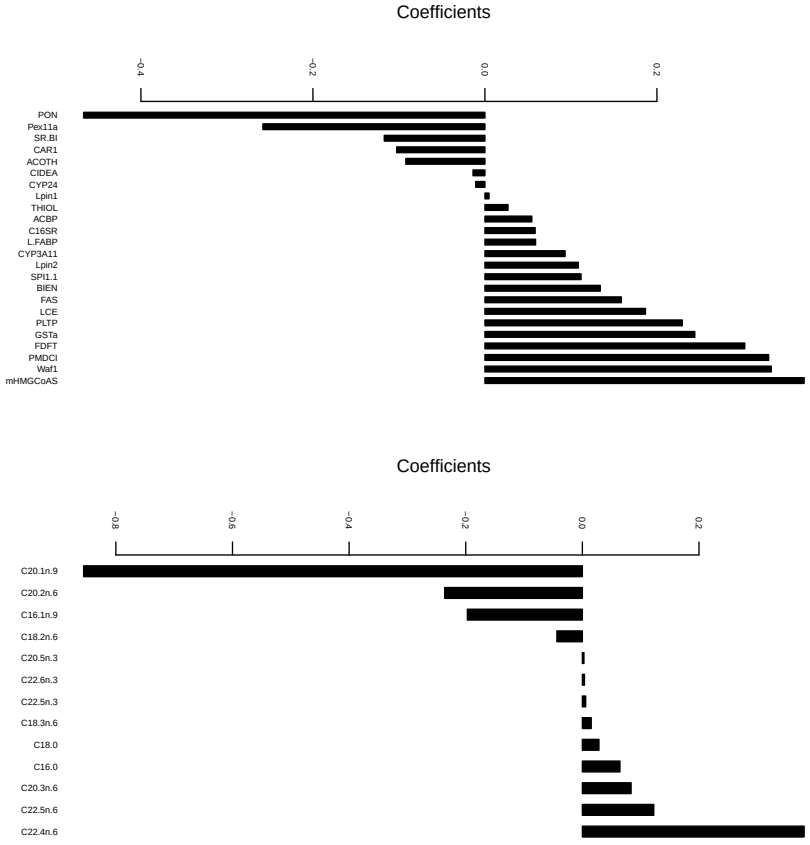

Figure 2: Nutrimouse data set: Coefficients of selected genes (top) and coefficients of selected fatty acids (bottom) in the first canonical vector pair.

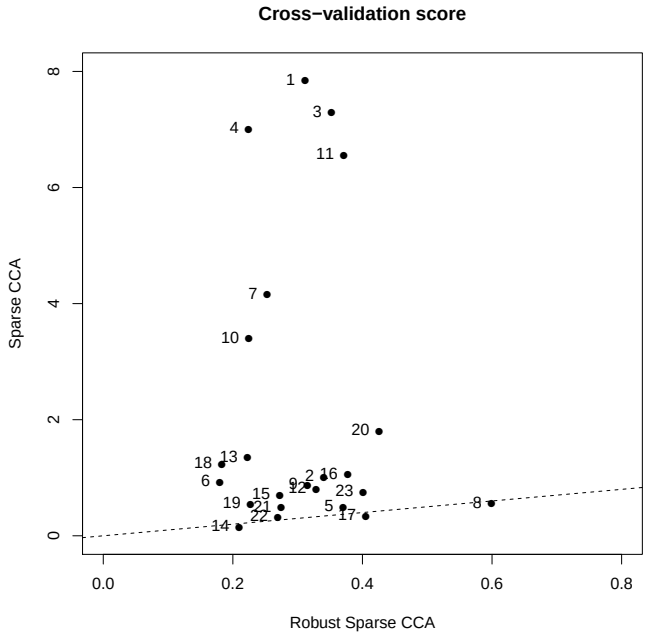

Figure 3: Breast cancer data set: 23 cross-validation scores (one for each chromosome) for Robust Sparse CCA (horizontal axis) and Sparse CCA (vertical axis). The dashed line is the 45°-line.

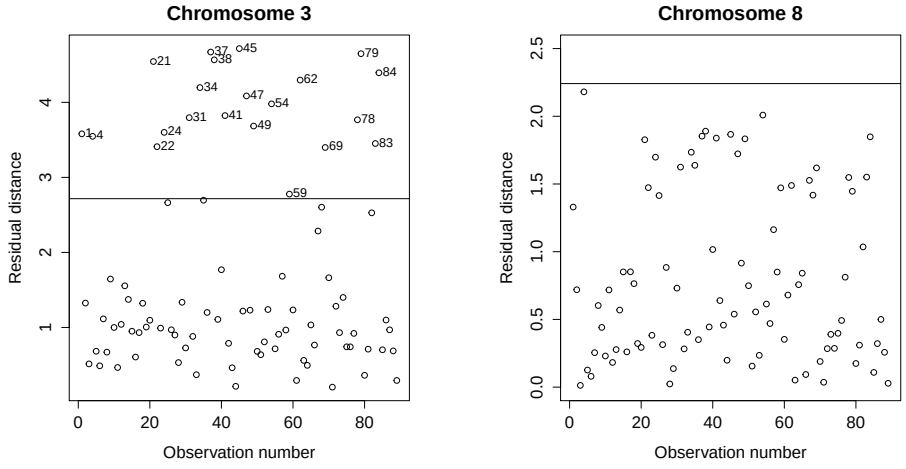

Figure 4: Breast cancer data set: Residual Distance Plot for chromosome 3 (left) and chromosome 8 (right).

Tables

Table 1: Simulation designs.

Design	$\mathbf{\Sigma}_{xx}$	$\mathbf{\Sigma}_{yy}$	$\mathbf{\Sigma}_{xy}$
Uncorrelated Sparse Low-dimensional $n = 100, p = 6, q = 4$	$10^{-2} \cdot \mathbf{I}_p$	$10^{-2} \cdot \mathbf{I}_q$	$10^{-2} \cdot \begin{bmatrix} 0.9 & \mathbf{0}_{1\times 3} \\ \mathbf{0}_{5\times 1} & \mathbf{0}_{5\times 3} \end{bmatrix}$
Correlated Sparse Low-dimensional $n = 100, p = 6, q = 4$	$10^{-2} \cdot \begin{bmatrix} 1 & 0.4 & \mathbf{0} \\ 0.4 & 1 & \mathbf{0} \\ 0 & 0 & \mathbf{I}_{4\times 4} \end{bmatrix}$	$10^{-2} \cdot \begin{bmatrix} 1 & 0.4 & \mathbf{0} \\ 0.4 & 1 & \mathbf{0} \\ 0 & 0 & \mathbf{I}_{2\times 2} \end{bmatrix}$	$10^{-2} \cdot \begin{bmatrix} 0.8 & \mathbf{0}_{1\times 3} \\ \mathbf{0}_{5\times 1} & \mathbf{0}_{5\times 3} \end{bmatrix}$
NonSparse Low-dimensional $n = 100, p = 12, q = 8$	$10^{-2} \cdot \mathbf{I}_p$	$10^{-2} \cdot \mathbf{I}_q$	$10^{-2} \cdot \mathbf{0.1}_{p\times q}$
Sparse High-dimensional 1 $n = 100, p = 100, q = 4$	$10^{-1} \cdot \mathbf{I}_p$	$10^{-1} \cdot \mathbf{I}_q$	$10^{-1} \cdot \begin{bmatrix} \mathbf{0.45}_{2\times 2} & \mathbf{0}_{2\times 2} \\ \mathbf{0}_{98\times 2} & \mathbf{0}_{98\times 2} \end{bmatrix}$
Sparse High-dimensional 2 $n = 50, p = q = 100$	$10^{-7} \cdot \begin{bmatrix} \mathbf{S}_{10\times 10} & \mathbf{0} \\ \mathbf{0} & 10^{-3} \cdot \mathbf{I}_{90\times 90} \end{bmatrix}$ with $\mathbf{S}_{ij} = \begin{cases} 1 & if\ i = j \\ 0.8 & if\ i \neq j, \end{cases}$	$\mathbf{\Sigma}_{xx}$	$10^{-7} \cdot \begin{bmatrix} \mathbf{0.8}_{10\times 10} & \mathbf{0}_{10\times 90} \\ \mathbf{0}_{90\times 10} & \mathbf{0}_{90\times 90} \end{bmatrix}$
Sparse Ultra High-dimensional $n = 100, p = q = 10000$	$10^{-7} \cdot \begin{bmatrix} \mathbf{S}_{10\times 10} & \mathbf{0} \\ \mathbf{0} & 10^{-3} \cdot \mathbf{I}_{9990\times 9990} \end{bmatrix}$ with $\mathbf{S}_{ij} = \begin{cases} 1 & if\ i = j \\ 0.8 & if\ i \neq j, \end{cases}$	$\mathbf{\Sigma}_{xx}$	$10^{-7} \cdot \begin{bmatrix} \mathbf{0.8}_{10\times 10} & \mathbf{0}_{10\times 9990} \\ \mathbf{0}_{9990\times 10} & \mathbf{0}_{9990\times 9990} \end{bmatrix}$

Table 2: Simulation results. Average of the angles between the space spanned by the true and estimated canonical vectors; average true positive rate and true negative rate are reported for each method.

Design	Method	No contamination			t-distribution			Contamination		
		$\bar{\theta}(\hat{\mathbf{A}}, \mathbf{A})$	TPR	TNR	$\bar{\theta}(\hat{\mathbf{A}}, \mathbf{A})$	TPR	TNR	$\bar{\theta}(\hat{\mathbf{A}}, \mathbf{A})$	TPR	TNR
Uncorrelated	CCA	0.11	1.00	0.00	0.22	1.00	0.00	0.38	1.00	0.00
Sparse	Robust CCA	0.14	1.00	0.00	0.15	1.00	0.00	0.15	1.00	0.00
Low-dimensional	Sparse CCA	0.04	0.98	0.97	0.19	0.94	0.63	0.34	1.00	0.04
	Robust Sparse CCA	0.04	1.00	0.82	0.11	1.00	0.52	0.05	1.00	0.76
Correlated	CCA	0.06	1.00	0.00	0.13	1.00	0.00	0.43	1.00	0.00
Sparse	Robust CCA	0.08	1.00	0.00	0.09	1.00	0.00	0.09	1.00	0.00
Low-dimensional	Sparse CCA	0.13	1.00	1.00	0.19	0.96	0.76	0.57	0.52	0.02
	Robust Sparse CCA	0.07	1.00	0.57	0.09	1.00	0.34	0.07	1.00	0.53
NonSparse	CCA	0.08	1.00	NA	0.32	1.00	NA	0.20	1.00	NA
Low-dimensional	Robust CCA	0.11	1.00	NA	0.12	1.00	NA	0.12	1.00	NA
	Sparse CCA	0.41	0.93	NA	0.67	0.82	NA	0.23	1.00	NA
	Robust Sparse CCA	0.16	0.99	NA	0.22	0.99	NA	0.13	1.00	NA
Sparse	Sparse CCA	0.65	0.62	0.99	0.70	0.71	0.87	0.36	1.00	0.80
High-Dimensional 1	Robust Sparse CCA	0.66	0.84	0.86	0.56	0.82	0.86	0.16	0.96	0.97
Sparse	Sparse CCA	1.08	0.31	1.00	1.14	0.23	1.00	1.25	0.38	0.97
High-Dimensional 2	Robust Sparse CCA	0.59	0.87	0.87	0.60	0.94	0.89	0.84	0.97	0.82
Sparse Ultra	Sparse CCA	1.18	0.17	1.00	1.22	0.15	1.00	1.25	0.40	1.00
High-dimensional	Robust Sparse CCA	1.42	0.93	1.00	1.24	0.98	1.00	0.98	1.00	1.00

Table 3: As in Table 2, comparing Robust Sparse CCA to other alternatives in the "Sparse High-dimensional 2 design".

Method	No contamination			t-distribution			Contamination		
	$\bar{\theta}(\hat{\mathbf{A}}, \mathbf{A})$	TPR	TNR	$\bar{\theta}(\hat{\mathbf{A}}, \mathbf{A})$	TPR	TNR	$\bar{\theta}(\hat{\mathbf{A}}, \mathbf{A})$	TPR	TNR
Sparse CCA of Parkhomenko et al. (2009)	0.93	1.00	0.93	1.41	0.94	0.72	1.28	0.89	0.00
Sparse CCA of Witten et al. (2009)	0.79	0.65	1.00	1.16	0.30	0.92	1.57	0.00	0.00
Sparse CCA of Waaijenborg et al. (2008)	0.44	1.00	0.08	1.01	1.00	0.02	1.25	1.00	0.00
Sparse CCA on pre-processed data	0.58	0.92	0.79	0.72	0.88	0.74	1.36	0.74	0.25
Sparse CCA with robust initialization	1.07	0.32	1.00	1.13	0.24	1.00	1.25	0.38	0.97
Robust Sparse CCA	0.59	0.87	0.87	0.60	0.94	0.89	0.84	0.97	0.82

Table 4: Evaporation data set: Cross-validation score for standard CCA, Robust CCA, Sparse CCA and Robust Sparse CCA.

Method	CV-score 0% Trimming	CV-score 10% Trimming
CCA	0.74	0.49
Robust CCA	0.57	0.39
Sparse CCA	0.57	0.41
Robust Sparse CCA	0.48	0.31

Table 5: Evaporation data set: Estimated canonical vectors using Robust CCA and Robust Sparse CCA.

		Robust CCA		Robust Sparse CCA	
	Variables \ Canonical Vectors	1	2	1	2
First	MAXST: Max. daily soil temperature	-0.35	-0.76	0	-0.70
data	MINST: Min. daily soil temperature	0.03	0.63	0	0.71
set	AVST: Avg. daily soil temperature	0.93	0.18	1	0
Second	MAXAT: Max. daily air temperature	0.54	-0.11	0.94	0
data	MINAT: Min. daily air temperature	0.67	0.84	0.14	0.38
set	AVAT: Avg. daily air temperature	0.14	-0.03	0.17	0.36
	MAXH: Max. daily relative humidity	-0.13	0.09	0	0
	MINH: Min. daily relative humidity	-0.03	0.36	0	0.85
	AVH: Avg. daily relative humidity	-0.28	0.32	-0.24	0
	WIND: Total wind, measured in miles per day	-0.37	-0.19	0	0
	Canonical correlations	0.93	0.56	0.87	0.48

Table 6: Nutrimouse data set: Cross-validation score for Sparse CCA and Robust Sparse CCA.

Method	CV-score 0% Trimming	CV-score 10% Trimming
Sparse CCA	98.78	92.53
Robust Sparse CCA	6.30	4.31